Chinese restaurant process from stick-breaking for Pitman-Yor

Caroline Lawless, Julyan Arbel

Chinese restaurant process from stick-breaking for Pitman–Yor

CAROLINE LAWLESS AND JULYAN ARBEL

UNIV. GRENOBLE ALPES, INRIA, CNRS, LJK, 38000 GRENOBLE, FRANCE

Introduction

- The Chinese restaurant process and the stick-breaking process are the two most commonly used representations of the Pitman–Yor process.
- However, the usual proof of the connection between them is indirect.
- Miller (2018) proved directly that the stick-breaking process gives rise to the Chinese restaurant process representation of the Dirichlet process.
- The Dirichlet process is a special case of the Pitman–Yor process.
- We extend Miller's proof to Pitman–Yor process random measures.

Pitman–Yor & Dirichlet processes

- The Dirichlet Process (DP) and the Pitman–Yor process (PY, Pitman and Yor, 1997) are discrete random probability measures.
- The PY is parametrized by $d \in (0,1), \alpha > -d$, and a base probability measure P_0. The DP is recovered by letting $d = 0$.
- The stick-breaking representation (Sethuraman, 1994) is given by

$$v_i \sim \begin{cases} \text{Beta}(1,\alpha) & \text{for DP} \\ \text{Beta}(1+d, \alpha+id) & \text{for PY} \end{cases}$$

$$\pi_k = v_k \prod_{i=1}^{k-1}(1-v_i),\ \phi_k \overset{\text{iid}}{\sim} P_0.$$

We define the random process P by

$$P = \sum_{i=1}^{\infty} \pi_k \delta_{\phi_k}.$$

- The Chinese restaurant process (Antoniak, 1974) is the distribution induced on random partitions $\boldsymbol{C}$ given by

$$P(\boldsymbol{C} = C) = \begin{cases} \frac{\alpha^{|C|}\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c\in C} \Gamma(|\,c\,|) & \text{for DP} \\ \frac{d^t (\frac{\alpha}{d})_{(t)}}{(\alpha)_{(n)}} \prod_{j=1}^{t} (1-d)_{(|c_j|-1)} & \text{for PY.} \end{cases}$$

Theorem

Suppose $\boldsymbol{\pi}$ follows the PY stick-breaking, and

$$\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n | \boldsymbol{\pi} = \pi \overset{\text{iid}}{\sim} \pi, \text{ that is, } \mathbb{P}(\boldsymbol{z}_i = k \mid \pi) = \pi_k,$$

and $\boldsymbol{C}$ is the partition of $[n]$ induced by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$.
Then $\boldsymbol{C}$ follows the PY Chineses restaurant process.

Technical Lemmas

Our proof relies on the following lemmas, which here we will state without proof. Let us abbreviate $z = (z_1, \ldots, z_n)$. Given $z \in \mathbb{N}^n$, let C_z denote the partition $[n]$ induced by z. We define $m(z) = \max\{z_1, \cdots, z_n\}$, and $g_k(z) = \#\{i\colon z_i \geq k\}$.

Lemma 1 For any $z \in \mathbb{N}^n$,

$$\mathbb{P}(\boldsymbol{z} = z) = \frac{1}{(\alpha)_{(n)}} \prod_{c\in C_z} \frac{\Gamma(|c|+1-d)}{\Gamma(1-d)} \prod_{k=1}^{m(z)} \frac{\alpha+(k-1)d}{g_k(z)+\alpha+(k-1)d}.$$

Lemma 2 For any partition C of $[n]$,

$$\sum_{z\in\mathbb{N}^n} \mathbb{1}(C_z = C) \prod_{k=1}^{m(z)} \frac{\alpha+(k-1)d}{g_k(z)+\alpha+(k-1)d} = \frac{d^t(\frac{\alpha}{d})_{(t)}}{\prod_{c\in C}(|c|-d)}.$$

Proof of Theorem

$$\mathbb{P}(\boldsymbol{C} = C) = \sum_{z\in\mathbb{N}^n} \mathbb{P}(\boldsymbol{C} = C | \boldsymbol{z} = z)\mathbb{P}(\boldsymbol{z} = z)$$

$$\overset{(a)}{=} \sum_{z\in\mathbb{N}^n} \mathbb{1}(C_z = C) \frac{1}{(\alpha)_{(n)}} \prod_{c\in C_z} \frac{\Gamma(|c|+1-d)}{\Gamma(1-d)} \prod_{k=1}^{m(z)} \frac{\alpha+(k-1)d}{g_k(z)+\alpha+(k-1)d}$$

$$= \frac{1}{(\alpha)_{(n)}} \prod_{c\in C_z} \frac{\Gamma(|c|+1-d)}{\Gamma(1-d)} \sum_{z\in\mathbb{N}^n} \mathbb{1}(C_z = C) \prod_{k=1}^{m(z)} \frac{\alpha+(k-1)d}{g_k(z)+\alpha+(k-1)d}$$

$$\overset{(b)}{=} \frac{1}{(\alpha)_{(n)}} \prod_{c\in C_z} \frac{\Gamma(|c|+1-d)}{\Gamma(1-d)} \frac{d^t(\frac{\alpha}{d})_{(t)}}{\prod_{c\in C}(|c|-d)}$$

$$\overset{(c)}{=} \frac{1}{(\alpha)_{(n)}} \prod_{c\in C} (1-d)_{(|c|-1)} \prod_{c\in C} (|c|-d) \frac{d^t(\frac{\alpha}{d})_{(t)}}{\prod_{c\in C}(|c|-d)}$$

$$= \frac{d^t(\frac{\alpha}{d})_{(t)}}{(\alpha)_{(n)}} \prod_{j=1}^{t} (1-d)_{(|c_j|-1)}$$

where (a) is by Lemma 1, (b) is by Lemma 2, and (c) is since $\Gamma(|c|+1-d) = (|c|-d)\Gamma(|c-d|)$.

Further research

- The Dirichlet process and the Pitman–Yor process are only special cases of a broad class of random measures called Gibbs-type random measures.
- An interesting further study would be to investigate the possibility of extending this proof to Gibbs-type random measures.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.

Miller, J. W. (2018). An elementary derivation of the Chinese restaurant process from Sethuraman's stick-breaking process. *arXiv preprint arXiv:1801.00513*.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.